With the development of urban environments, it is imperative that policy makers allocate resources towards public works projects for the benefit of the community. However, such policy makers may be incentivized to localize the benefit that the "community" receives to a particular area. Ironically and unfortunately, these resources tend to go to those who need it least in order to attract wealthy people to certain areas and localize those with lower incomes to higher crime areas.

Thus, these data sets can be combined to show how the development of public infrastructure and utilities, such as access to the MBTA, open spaces, and community pools can have a direct impact on the property values of an area as well as it's crime rate. Such an analysis can bring to light the issue at hand to the citizens and policy makers that may be unaware of such issues. Perhaps it is even possible, with such a data driven analysis to influence data driven decision making to determine where to develop future projects in order to maximize its utilitarian impact per dollar amount. Indeed, such a future would certainly align with a utopic view of a truly smart city.

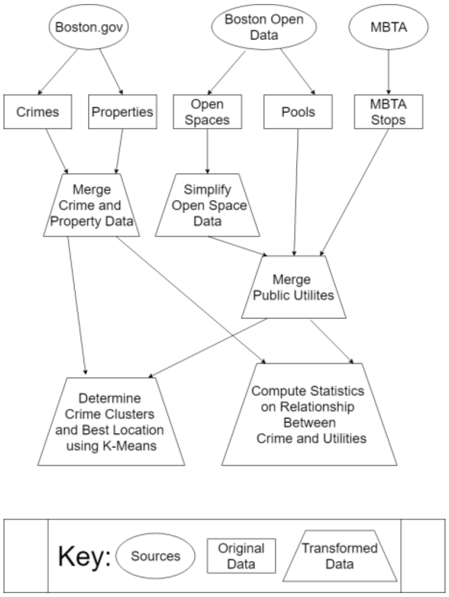
The data for crime and property data was pulled from https://data.boston.gov in crime_data.py and property_data.py. The data for open space and pool data was pulled from http://bostonopendata-boston.opendata.arcgis.com in open_space_data.py and pool_data.py. Finally, the MBTA data was pulled from https://api-v3.mbta.com in mbta_stop_data.py.

For my first transformation, simplify_open_space.py, I decided to simplify the open space data in simplify_open_space.py to give it a non polygonal coordinate as an estimate to allow it to more easily interact with the other data sets. I did this by taking the average of the longitudes and latitudes for each polygon and use that as a center representation of the open space. I then calculated a rough circumradius by assuming the polygons where roughly normal in order to provide a very rough estimation as to the size of such parks.

For my second transformation, combine_public_utilities.py, I decided to merge all of the public utilities into one format in order to have one source of longitude and latitude data, as well as extra information about each of the utilities.

For my third transformation, combine_crime_and_property.py, I merged the crime and property data to create a collection that can represent which types of streets had what types of crimes occured on them and their associated property values.

The outputs of the transformations are found in the json files: open_spaces_simplified.json, public_utilities.json, and crime_and_properties.json. A shorthand notation of the methods I used to generate these datasets are shown below:

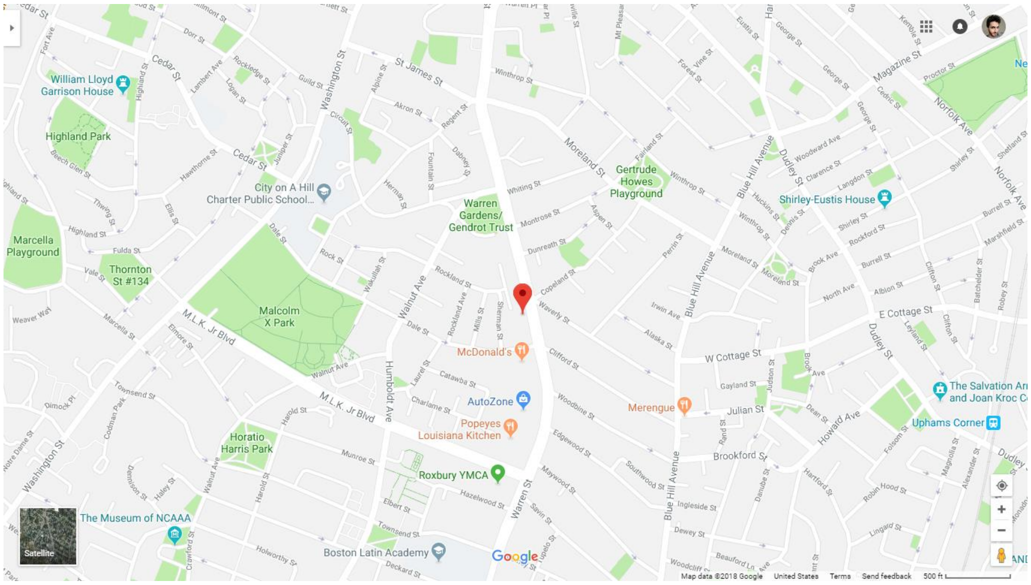


After collecting this data, I became interested in analyzing how crime is affected by public services in Boston. In theory, creating more public utilities should decrease the associated crime in the area as there would be cheap, fun, and legal ways to occupy time instead of resorting to crime.

To test this hypothesis, I employed constraint satisfaction using the k-means algorithm in get_crime_clusters.py to determine at most three separate locations for new public utilities that would have the maximum impact at reducing crime. In order to do this, I theorized that if a new public service were to be placed, it should be placed in an area where crime rates are higher and where there are no already existing public works within a 5 minute walk (roughly equivalent to (80m/min)*5 min = 400 m according to https://en.wikipedia.org/wiki/Walking_distance_measure). Thus, the k-means algorithm would determine certain crime areas with abnormally high crime zones and removed the zones that did not satisfy my extra constraints.

While the k-means algorithm served to solve the problem at hand, the runtime was very inefficient, and would only run within a reasonable amount of time with a subset of my data. Thus, I ran my k-means algorithm in “trial mode” and only used 999 crime locations in Boston to find my clusters.

After running crime_clusters.py in trial mode, and after filtering, there was only one point which satisfied the constraints. This point is located near Dudley Square in Roxbury and the nearest public service in my dataset to the point, Warren Gardens/Gendrot Trust, is 486.2m away. Indeed, this could serve as a potential location to develop a new public service to hopefully reduce crime. This location is shown on the map below:



After determining a crime center that fit my parameters, I wanted to see what types of public utilities should in theory produce the best outcome when it comes to decreasing crime. In order to do this, I decided to find the correlation between distance to the three types of public utilities in my data, (MBTA Stops, Pools, Open Spaces) and the amount of crimes that occur using utilities_vs_crime.py.

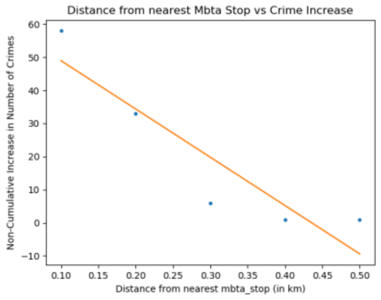
Here are the correlations I calculated from my analysis:

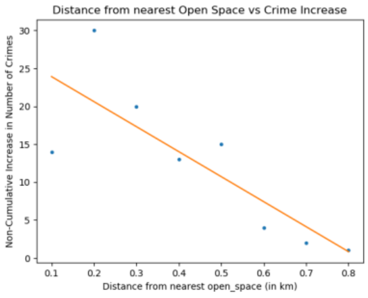
mbta_stop: -0.9177490484590687

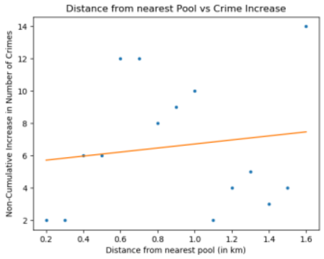
pool: 0.13925778432983166

open_space: -0.8160234544734344

Below are the corresponding graphs generated, and the best fit lines for each:







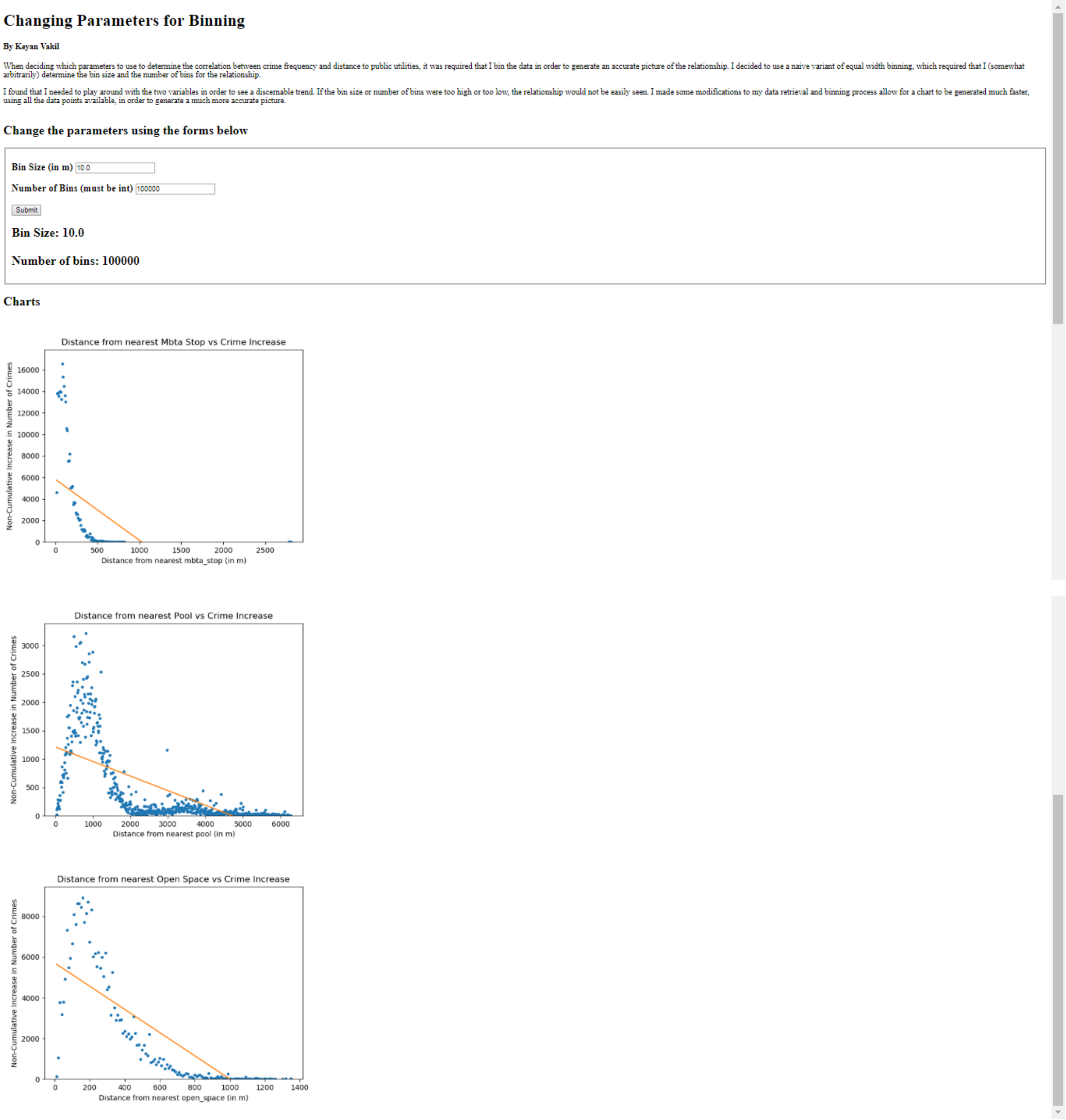
There are certainly surprising results here. I predicted that as the distance to public services increase, the number of crimes would also increase. However, it seems that closer proximity to MBTA stops and open spaces is correlated to more crime. Additionally, open problems still remain. Do these correlations reveal the problem at hand, or is there another confounding variable such as the increase in traffic that contributes to the discrepancy of my hypothesis to the results.

A major assumption I made that may explain these results is that I assumed that the frequency of people located around these areas were constant. In other words, this is a measure of proximity to public services and the number of crimes that occur, not necessarily the crime rate, which would perhaps portray a different result. The only positively correlated result was pools, but even then, the correlation is surprisingly quite low.

The proposed problem with my analysis can be ameliorated by normalizing the crime rate data with foot traffic data, so that we know we can assess the relative crime rate of the area as opposed to the absolute crime rate. Additionally, future work can take into account historical data to see whether the placement of such public services actually led to a decrease in the crime rate.

After generating the graphs, I discovered that the binning technique I was using to determine the crime increases as the distance increased from a public utility was somewhat arbitrary. I was using a binning technique that possessed bins of equal width, but did not consider a min or max value in my binning, which would allow for a less arbitrary equal width binning. As a result, I created a small Flask web application in the Project3 directory such that one can change the size of each bin and the number of bins, which would result in different graphical visualizations of the statistical data.

In order to ensure a reasonable runtime, I preprocessed some data using get_min_dists.py to generate min_dists.csv, which hold a csv where each line represent a crime and each value represent the nearest distance from the corresponding public utility in kilometers. The flask application depends on min_dists.csv to run properly. To test it out yourself, ensure that the you install the requirements listed in requirements.txt into your environment, set the “FLASK_APP” environment variable to Project3/app.py then run `flask run` to load a local server. A picture of the resulting application is shown below.



Changing Parameters for Binning

By Keyan Vakil

When deciding which parameters to use to determine the correlation between crime frequency and distance to public utilities, it was required that I bin the data in order to generate an accurate picture of the relationship. I decided to use a naive variant of equal width binning, which required that I (somewhat arbitrarily) determine the bin size and the number of bins for the relationship.

I found that I needed to play around with the two variables in order to see a discernable trend. If the bin size or number of bins were too high or too low, the relationship would not be easily seen. I made some modifications to my data retrieval and binning process allow for a chart to be generated much faster, using all the data points available, in order to generate a much more accurate picture.

Change the parameters using the forms below

Bin Size (in m) 10.0

Number of Bins (must be int) 100000

Bin Size: 10.0

Number of bins: 100000

Charts

Distance from nearest Mbta Stop vs Crime Increase
Non-Cumulative Increase in Number of Crimes
0
2000
4000
6000
8000
10000
12000
14000
16000
0
500
1000
1500
2000
2500
Distance from nearest mbta_stop (in m)
Distance from nearest Pool vs Crime Increase
Non-Cumulative Increase in Number of Crimes
0
500
1000
1500
2000
2500
3000
0
1000
2000
3000
4000
5000
6000
Distance from nearest pool (in m)
Distance from nearest Open Space vs Crime Increase
Non-Cumulative Increase in Number of Crimes
0
2000
4000
6000
8000
0
200
400
600
800
1000
1200
1400
Distance from nearest open_space (in m)